REPORT DOCUMENTATION PAGE

AFRL-SR-AR-TR-05-

01.72

1. AGENCY USE ONLY *(Leave blank)*	**2. REPORT DATE** 31 Mar 05	**3.** Final Report 01 Oct 04 to 31 Mar 05

4. TITLE AND SUBTITLE
(STTR FY03) Application of Cortical Processing Theory to Accoustical Analysis

5. FUNDING NUMBERS
F49620-03-C-0051

6. AUTHOR(S)
Ghitza, Oded
Messing
Braida

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Sensimetrics Corporation
48 Grove Street - Suite 305
Somerville, MA 02144-2500

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Dr. Willard Larkin
AFOSR/NL
875 North Randolph Street
Suite 325, Room 3112
Arlington, VA 22203

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION AVAILABILITY STATEMENT
Approve for Public Release: Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

The overall goal of the STTR program is to formulate a template-matching operation, with perception-related rules of integration over time and frequency at its core, in the context of human perception of degraded speech. In particular, we aim at developing models of auditory processing capable of predicting consonant confusion by noramlly-hearing listeners, under a variety of acoustic distortions A prerequisite is to formulate the signal processing principles realized by the auditory periphery in providing the observed graceful degredation of human performance in noise.

In the nominal four quarters of Phase I we have focused on the role of the descending auditory pathway in regulating the operating point of the cochlea, resulting in auditory nerve (AN) representation of speech sounds that are less sensitive to changes in sustained background noise. A closed-loop model of the auditory periphery, with efferent-inspired feedback, has been implemented that produces spectrograms of noisy speech that are more consistent with spectrograms of speech in quiet than are spectrograms produced by open-loop models of the auditory periphery.

14. SUBJECT TERMS

15. NUMBER OF PAGES
23

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	**18. SECURITY CLASSIFICATION OF THIS PAGE**	**19. SECURITY CLASSIFICATION OF ABSTRACT**	**20. LIMITATION OF ABSTRACT**

<u>Final Report for STTR Contract No. F49620-03-C-0051, Item 0001AF (March 31, 2005)</u>

A. Introduction

This is a final report for an extended Phase I of the STTR program entitled "Application of Cortical Processing Theory to Acoustical Analysis". Two quarters were added to the nominal Phase I four quarters (10/01/2003 – 09/31/2004), covering the time-period 10/1/2004 – 03/31/2005. A final report of the nominal Phase I was submitted on 09/31/2004 as item 0001AD and is reprinted here in Appendix A. The current report describes the progress made in the extended time period.

The overall goal of the STTR program is to formulate a template-matching operation, with perception-related rules of integration over time and frequency at its core, in the context of human perception of degraded speech. In particular, we aim at developing models of auditory processing capable of predicting consonant confusions by normally-hearing listeners, under a variety of acoustic distortions. A prerequisite is to formulate the signal processing principles realized by the auditory periphery in providing the observed graceful degradation of human performance in noise.

In the nominal four quarters of Phase I we have focused on the role of the descending auditory pathway in regulating the operating point of the cochlea, resulting in auditory nerve (AN) representation of speech sounds that are less sensitive to changes in sustained background noise. A closed-loop model of the auditory periphery, with efferent-inspired feedback, has been implemented that produces spectrograms of noisy speech that are more consistent with spectrograms of speech in quiet than are spectrograms produced by open-loop models of the auditory periphery. The model is described in Appendix A.

B. Towards tuning the Peripheral Auditory Model (PAM)

A need arises for a quantitative methodology to evaluate the adequacy of the PAM to preserve phonetic information that is perceptually relevant. The challenge is to measure the errors due to the PAM <u>in isolation</u> from errors created by the "back-end" of the evaluation tool. The approach we are undertaking is to develop a methodology that brings errors due to the back-end closer to zero. Towards this end we are using the DRT paradigm (Voiers, 1983), briefly overviewed in Appendix B. Inspired by Hant and Alwan (2003)[1] we use "frozen speech" stimuli, namely, the same acoustic token is being used for training and for testing, hence the testing token differs from the training token only by the acoustic distortion. In the frozen-speech methodology we also assume that the acoustic realizations of the final diphone in both stimuli of a DRT pair are identical, hence resulting in zero-error contributions to the L_2-norm distance measure used by the DRT mimic. This, however, is not the case even though we restricted the database to utterances spoken by one male speaker pronouncing the utterances very carefully.

Consequently, to further reduce errors due to the back-end, we are modifying the frozen-speech methodology by considering an acoustic realization of the DRT word-pairs produced by a speech synthesizer. The goal is to generate speech stimuli such that, for a given word-pair, the

[1] Hant and Alwan (2003) evaluated the performance of a functional auditory model in predicting complex-signal discrimination in noise. Their tasks included discrimination of spectro-temporal patterns such as formant sweeps and synthetic CV syllables. Performance was measured for a discrimination task between two frozen stimuli (which in a detection task is 'noise' or 'signal-plus-noise') by making predictions based on <u>cell-by-cell differences</u> (in the L_2-norm sense) between the two stimuli, where a 'cell' is a small region in the time-frequency representation.

formants' target values of the vowel are identical, hence forcing zero-error contribution to the L_2-norm distance measured over time-frequency cells associated with the final diphone. We are using Sensimetrics' own speech synthesis product HLsyn[2], tuned to our purpose. Figure 1 demonstrates the extent to which we achieve this goal, for the word-pair *Joe-go*. Figs. 1(a) and 1(b) show displays of the simulated IHC responses after temporal smoothing for the words *Joe* and *go*, respectively, in quiet. Only the first 400ms are shown. Fig. 1(c) shows the absolute value of the difference between the displays on a pixel-by-pixel basis, and Fig. 1(d) shows the time evolution of the L_2-norm distance between the two representations, measured across frequency. In accordance with our goal, the L_2-norm distance outside the time interval spanned by the initial diphone approaches zero. Informal listening to the synthetic DRT database suggests a good speech quality; in quiet, DRT test results in a perfect score.

During the past quarter we have generated two sets of recordings in quiet and in noisy conditions, one for naturally spoken DRT words and one for synthetic DRT words. We used speech-shape noise at three intensities (70, 60 and 50 dbSPL) and at three SNRs (10, 5 and 0dB). Currently we conduct a formal DRT test for both sets, using 9 subjects with 4 repetitions each (this will set the variance to 2%). This data is needed in order to test whether the usage of synthetic stimuli worsens human performance (especially in the presence of noise); a significant drop in performance may indicate an increased role of the cognitive layers.

Figures 2 and 3 present DRT scores for naturally spoken (Fig. 2) and synthetic (Fig. 3) DRT words in the presence of speech-shape noise, with dbSPL (columns) and SNR (rows) as parameters. Shown are scores of one listener averaged over 4 repetitions. Overall number of errors ("Grand Mean") is somewhat higher for the synthetic DRT stimuli, but the error distribution among the phonetic features are reasonably similar. One exception in the Nasality dimension; Much fewer errors are being made listening to the synthetic stimuli. A thorough comparison will be made once the data from all 9 subjects will be collected.

Assuming a reasonable performance for the synthetic DRT stimuli we shall tune the efferent-inspired closed-loop auditory model using an iterative procedure. We shall adjust the parameters of the model with the goal of predicting the human response; adjusting the parameters will be constrained to processing principles that are plausible according to current understanding of the morphology and the neurophysiology of the peripheral auditory system (both ascending and descending pathways).

C. Towards a distance between diphones with different time scales

During this quarter we have begun a process of evaluating different methods to perform template matching. We shall be focusing on initial diphones because of the central role they play in spoken language. Linguistic studies show the relative importance of the initial diphone across many languages (e.g. Greenberg, 1978; Ohala, 1997), and psychophysical studies show its relative importance in phone perception (e.g. Ghitza, 1993).

Conceptually, the input to the template-matching box is an auditory representation of the acoustic signal. For diphone identification task we currently study a model suggested by Hopfield (2004). The model is motivated by processing principles that have been observed by

[2] The HLsyn speech synthesis system was developed by Stevens and colleagues at Sensimetrics Corporation (Stevens and Bickley, 1991; Hanson and Steven, 2002). It is a quasi-articulatory based system driven by a 13-component vector updated every 5ms. The 13 components are the fundamental frequency, the first four formants, and eight more parameters related to the physiology of speech production.

neurophysiological measurements in basic neural circuits in cortical areas. From our perspective, the most relevant property is the apparent invariance to time-scale in spatiotemporal input patterns. We have begun to study Hopfield's model in two fronts:

1. We have implemented a computational model following guidelines provided by Hopfield (in Matlab). The model uses crude front-end in the form of 20 channels, equally distributed on the MEL scale. Each channel feeds an array of 100 "Layer-I" Integrate-And-Fire (IAF) neurons. These neurons differ only in their threshold of firing; the thresholds are linearly distributed. All 2000 Layer-I neurons are synchronized via a weak-intensity low-frequency oscillation drive. For each MEL channel a "patch" of 10 successive neurons is formed, chosen at random from the 100 Layer-I neuron-array. The model also uses 2000 "Layer-II" IAF neurons; the input to a given Layer-II neuron is a linear combination (with fixed weights) of 5 Layer-I patches; the patches are chosen at random[3]. The input to a Layer-I neuron is the critical-band power. A differential equation is solved, with physiologically-plausible parameters. When the action-potential reaches the threshold a spike is fired, the action-potential is being reset to zero and a refractory-time is being set. The Layer-II neurons serve, in principle, as coincidence detectors; input is the linear sum of Layer-I neuron spikes produced in the 5 patches. A particular time-frequency "signature" leads to synchronous firing of Layer-I neurons, which leads to a firing of a Layer-II neuron. Weights are equal and fixed (i.e. no learning/training). The collective firing of the Layer-II neuron-array at time t_0, normalized to [0, 1], represents the probability of the input time-interval prior to t_0 to be the acoustic realization of a particular diphone. The properties of the model are being studied computationally for various speech stimuli; we will document our finding in future reports.
2. Mohan Sondhi has begun an analytical study of the response of an IAF neuron to dynamic input signals. Note the non-linear nature of the IAF circuit. In particular, we would like to understand how the neuron responds to inputs that differ only in time-scale.

[3] In principle we could exhausts all combinations of 5 (patches) out of 20 (channels); this would result in too many layer-II neurons. Instead, a linear combination of 5 patches, chosen at random, is assigned for 2000 neurons.

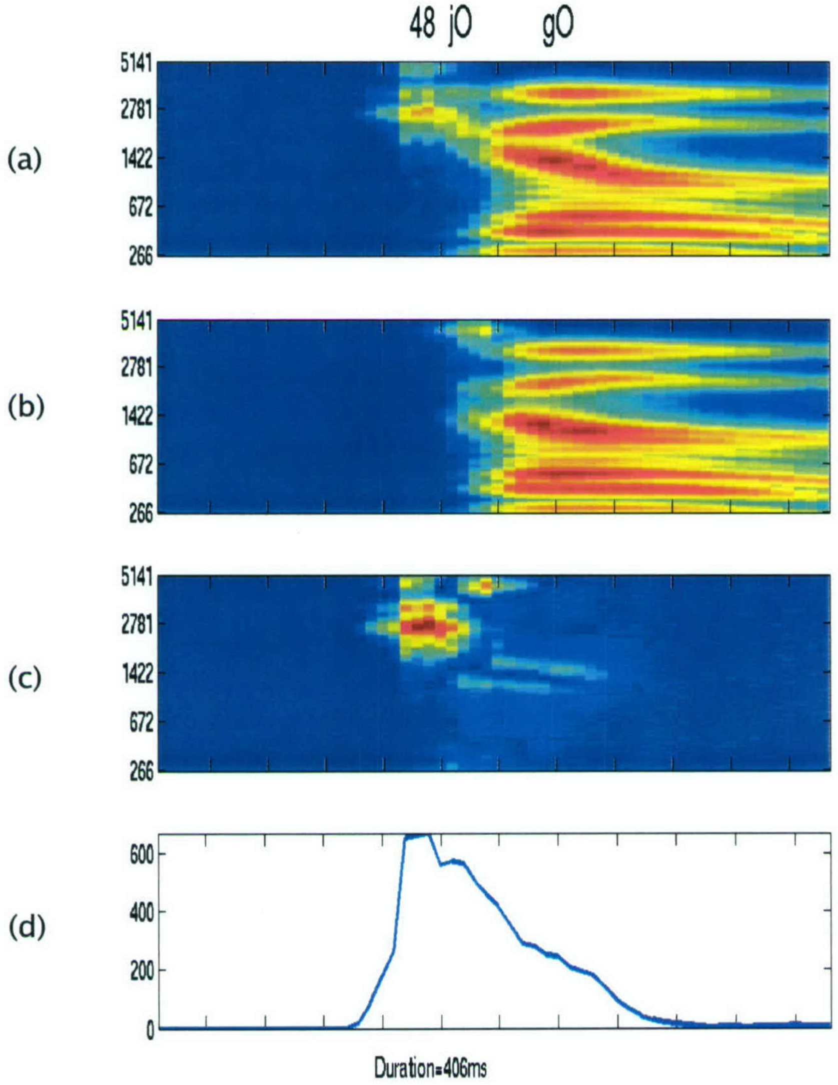

Figure 1. (a) "Rate" representation of the word *Joe* (first 400ms), in quiet. Abscissa - time; Ordinate - frequency. Consonant-to-vowel transition occurs at 200ms. (b) Same as (a) for the word *go*. (c) The absolute value of the difference between the displays on a pixel-by-pixel basis. (d) The time evolution of the L_2-norm distance between the two representations measured across frequency. The L_2-norm distance outside the time interval spanned by the initial diphone approaches zero.

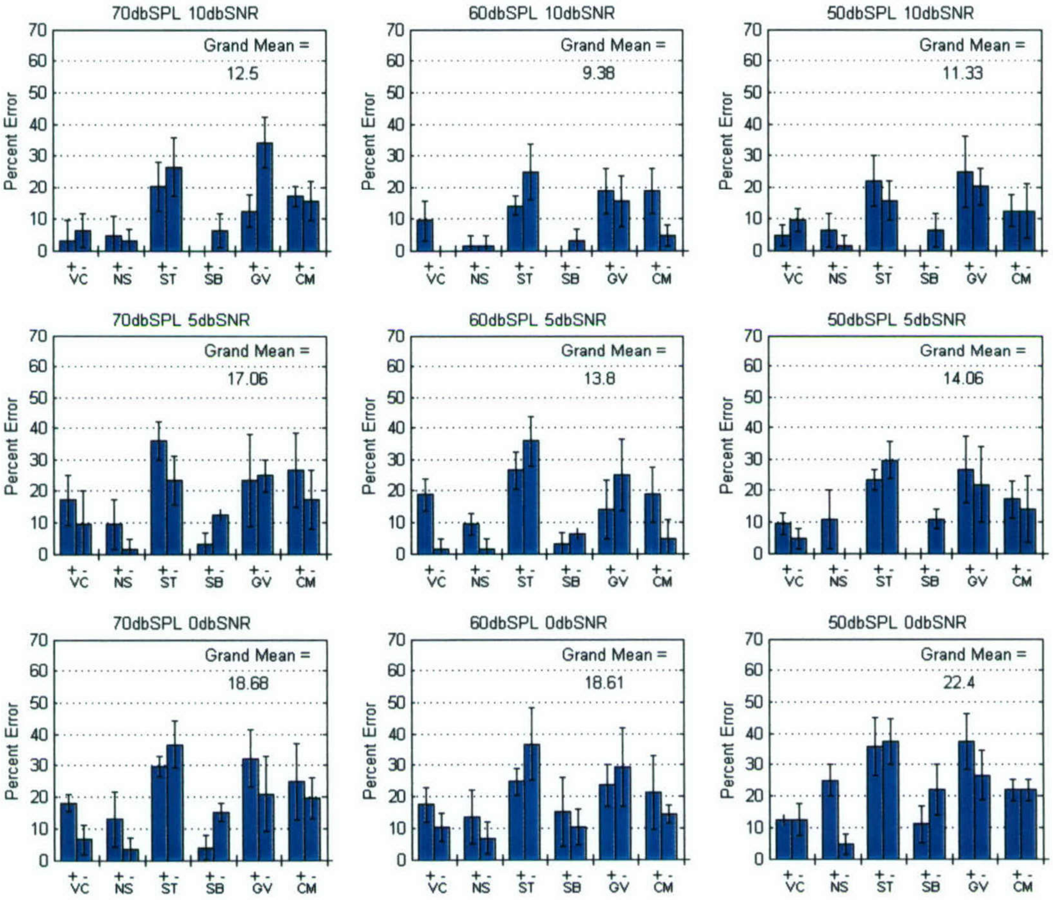

Figure 2. DRT scores for naturally spoken DRT words in the presence of speech-shape noise, with dbSPL (columns) and SNR (rows) as parameters. Shown are the resulting averages over four repetitions by one listener. The abscissa of every entry indicates the six phonemic categories: "VC" is for Voicing, "NS" for Nasality, "ST" for Sustention, "SB" for Sibilation, "GV" for Graveness and "CM" for Compactness. The "+" sign stands for attribute present and the "-" sign for attribute absent. The ordinate is termed "Error", and it represents the percentage of words in the category that, when played to the listener, were judged to be the opposite word in the word pair (i.e., the listener switched to the opposite category). Upper right corner indicates the Grand Mean (over all phonemic categories).

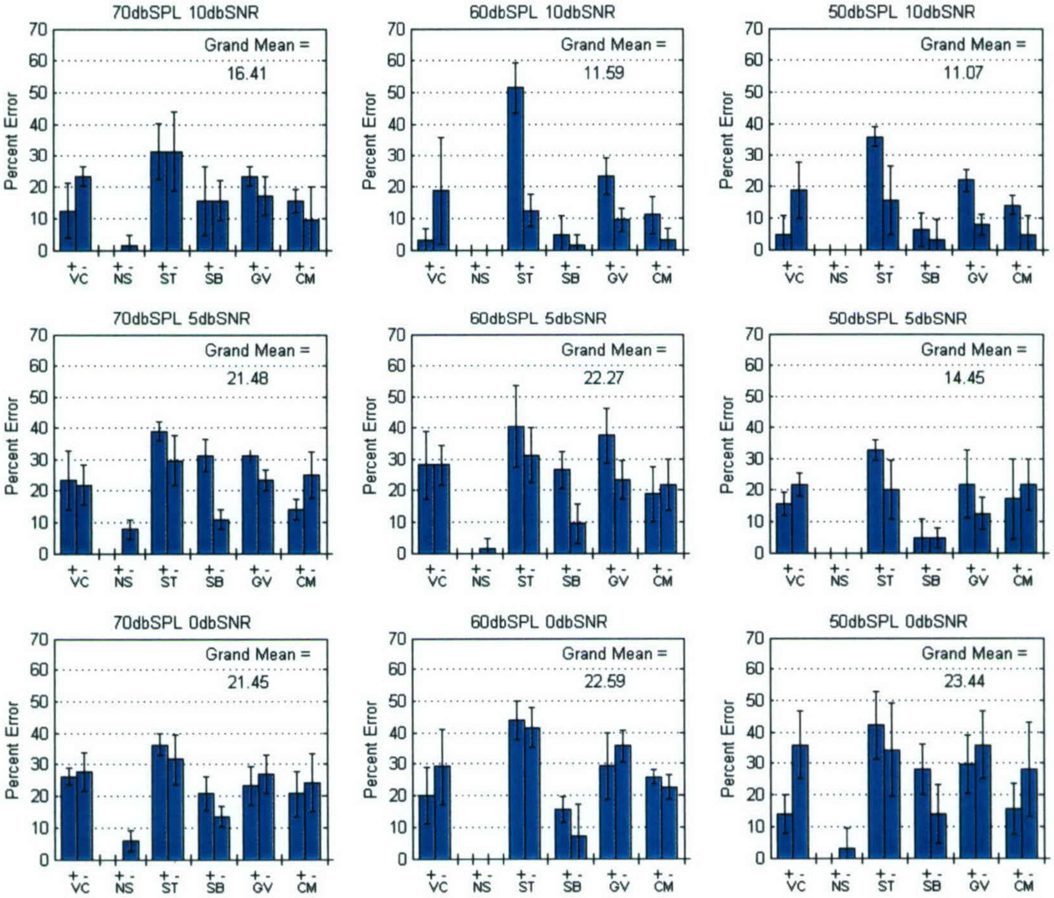

Figure 3. DRT scores for synthetic DRT words in the presence of speech-shaped noise. Conditions and legends are as in Fig. 2.

BIBLIOGRAPHY

Ghitza, O. (1993). Processing of spoken CVCs in the auditory periphery: I. Psychophysics. *J. Acoust. Soc. Am.*, 94(5), 2507-2516.

Greenberg, J. (1978). Some generalizations concerning initial and final consonants clusters. In J. Greenberg (ed.), *Universals of human language*. Stanford Ca: Stanford Univ. Press. (Volume 2: Phonology, 243-280).

Hanson, H.M. and K.N. Stevens (2002) A quasiarticulatory approach to controlling acoustic source parameters in a Klatt-type formant synthesizer using Hlsyn. *J. Acoust. Soc. Am.* 112, 1158-1182.

Hant, J.J., and Alwan, A. (2003). "A psychoacoustic-masking model to predict the perception of speech-like stimuli in noise," *Speech Communication*, 40, 291-313.

Hopfield, J. J. (2004). Encoding for computation: recognizing brief dynamical patterns by exploiting effects of weak rhythms on action-potential timing. *PNAS*, 101 (16) 6255-6260.

Ohala, J. J. and Kawasaki-Fukumori, H. (1997). Alternatives to the sonority hierarchy for explaining the shape of morphemes. In S. Eliasson and E. H. Jahr (eds.), *Studies for Einar Haugen*. Berlin: Mouton de Gruyter. 343-365.

Stevens, K.N. and Bickley, C.A. (1991). Constraints among parameters simplify control of Klatt formant synthesizer. *J. Phonetics* 19, 161-174.

Voiers, W. D. (1983). Evaluating processed speech using the diagnostic rhyme test. *Speech Technology*, 1(4):30–39.

Appendix A

Status Report for STTR Contract No. F49620-03-C-0051, Item 0001AD (Sept. 30, 2004)

A. INTRODUCTION

The overall goal of this STTR program is to formulate perception-related integration rules over time and frequency – presumably realized at post Auditory Nerve (AN) layers – in the context of speech perception in the presence of environmental noise. In particular, we aim at developing models of auditory processing capable of predicting phonetic confusions by normally-hearing listeners, under a variety of acoustic distortions. A prerequisite is to formulate the signal processing principles realized by the auditory system in providing the observed graceful degradation of human performance in noise.

Towards this end we suggest to model two interleaving functions: (1) the role of the descending pathway in regulating the operating point of the cochlea, resulting in AN representation of speech sounds that are less sensitive to changes in environmental conditions, and (2) the role of post-AN functions in extracting important acoustic-phonemic cues from the AN firing patterns. The underlying assumption is that the regulating mechanism and the post-AN mechanisms work in concert. Current models of the periphery are based upon the ascending pathway up through the AN. We propose to utilize the role of the descending pathway, mainly the Medial Olivocochlear (MOC[4]) feedback mechanism, and the way the ascending and the descending pathways interact. As a case study we shall focus on processing of speech in the presence of additive speech-shaped noise. It is suggested that the cochlear response in the presence of background noise is (much) more stable than the output from current feed-forward models. This observation is based upon the physiological and psychophysical evidence we currently have about the possible role of the MOC efferent system (see summary in Sec. B. of the report). To model functions of post-AN processing we propose a psychophysically based approach. The post-AN functions will be modeled as a template-matching system, where a time-frequency input pattern is matched against internal templates using a psychophysically derived distance measure. We suggest that the success of post-AN mechanisms in reliably extracting speech-related information in noise is partly due to the "stabilizing" effect of the efferent system.

This report summarizes work that has been completed in Phase I of the STTR program. We have implemented a closed-loop model of the auditory periphery with efferent-inspired feedback and have demonstrated its ability to produce spectrograms of noisy speech samples that are more consistent with spectrograms of speech in quiet than are spectrograms produced by open-loop models of the auditory periphery. As a baseline system we used a model of an open-loop linear cochlea whose details are described in Sec. C.1. In Sec. C.2. we compare the performance of the baseline system with that of a model of an open-loop nonlinear cochlea. In Sec. C.3 we introduce a model of closed-loop nonlinear cochlea with an efferent-inspired feedback.

The output of each model was defined as the temporal response of the simulated Inner Hair Cell (IHC) array, organized in the form of spectrograms. The output of the closed-loop model was compared quantitatively with the output of the baseline open-loop model. The criterion for comparison was the amount of consistency between the spectrographic representation of noisy

[4] The origin of the MOC nerve bundle is in the medial region of the superior olive, and it projects back to different places along the cochlea partition in a tonotopical manner, making synapse connections to the outer-hair cells. Detailed description is provided in Sec. B.

speech segments and that of the corresponding speech signals in quiet. Consistency was measured in terms of the distance between the noisy representations (with noise-intensity and SNR as parameters) and the representations of the speech in quiet (the reference). Sec. D. presents the quantitative evaluation. It shows that the closed-loop auditory model produces representations that are far more stable compared to those produced by the baseline (open-loop) auditory model. Whether this model of auditory periphery preserves phonetic information in patterns that follow psychophysical patterns will be rigorously inspected during Phase II, where the central part of the proposal, i.e. the formulation of a perception-based distance measure, will be established.

B. MOC EFFERENTS – BREIF REVIEW

B.1 MOC efferents: morphology and physiology

Numerous papers have been published providing detailed morphological and neurophysiological description of the medial olivocochlear (MOC) efferent feedback system (e.g., Guinan, 1996; May and Sachs, 1992; Winslow and Sachs, 1988). MOC efferents originate from neurons medial, ventral and anterior to the medial superior olivary nucleus (MSO), have myelinated axons, and terminate directly on Outer Hair Cells (OHC). Medial efferents project predominantly to the contralateral cochlea, the innervation is largest near the center of the cochlea, with the crossed innervation biased toward the base compared to the uncrossed innervation (e.g., Guinan, 1996). Roughly two-third of medial efferents respond to ipsilateral sound, one-third to contralateral sound, and a small fraction to sound in either ear. Medial efferents have tuning curves that are similar to, or slightly wider than, those of AN fibers, and they project to different places along the cochlear partition in a tonotopical manner. Finally, medial efferents have longer latencies and group delays than AN fibers. In response to tone or noise bursts, most MOC efferents have latencies of 10-40ms. Group delays measured from modulation transfer functions are much more tightly clustered, averaged at about 8ms. We currently do not have a clear understanding of the functional role of this mechanism. Few suggestions have been offered, such as shifting of sound-level functions to higher sound levels, antimasking effect on responses to transient sounds in a continuous masker, preventing damage due to intense sound (e.g., Guinan, 1996). One speculated role, which is of particular interest for this proposal, is a dynamic regulation of the cochlear operating point depending on background acoustic stimulation, resulting in robust human performance in perceiving speech in a noisy background. There are a few neurophysiologcal studies to support this role. Using anesthetized cats with noisy acoustic stimuli, Winslow and Sachs (1988), for example, showed that by stimulating the MOC nerve bundle electrically, the dynamic range of discharge rate at the AN is partly recovered. Measuring neural responses of awake cats to noisy acoustic stimuli, May and Sachs (1992) showed that the dynamic range of discharge rate at the AN level is only moderately affected by changes in levels of background noise.

B.2 MOC efferents: psychophysics – speech and speech-like stimuli

Few behavioral studies indicate the potential role of the MOC efferent system in perceiving speech in the presence of background noise. Dewson (1968) presented evidence that MOC lesions impair the abilities of monkeys to discriminate the vowel sounds [i] and [u] in the presence of masking noise but have no effect on the performance of this task in quiet. More recently, Giraud *et al*. (1997), and Zeng *et al*. (2000) showed that the performance of human subjects after they undergo a vestibular neurectomy (presumably resuling in a severed MOC feedback) deteriorates phoneme perception when the speech is presented in a noisy background. These speech reception experiments, however, provide questionable evidence because of surgical side effects such as uncertainties about the extent of the lesion and possible damage to cochlear elements. Recently, Ghitza (2004) quantified the role of the MOC efferent system by performing a test of initial consonant reception (the Diagnostic Rhyme Test) using subjects with normal hearing. Activation of selected parts of the efferent system was attempted by presenting speech and noise in various configurations (gated/continuous, monaural/binaural). Initial results of these experiments show a gated/continuous difference analogous to the 'masking overshoot' in tone detection. These results are interpreted to support the hypothesis of a significant efferent contribution to initial phone discrimination in noise.

B.3 Summary

Mounting physiological data exists in support of the effect of MOC efferents on the mechanical properties of the cochlea and, in turn, on the enhancement of signal properties at the auditory nerve level, in particular when the signal is embedded in noise. The current theory on the role of MOC efferents in hearing is that they cause a reduction in OHC motility and shape that results in increased basilar membrane stiffness which in turn produces an inhibited IHC response in the presence of noise that is comparable to the IHC response produced by a noiseless environment. We develop this popular theory into a closed-loop model of the peripheral auditory model that adaptively adjusts its cochlear operating point such that the time-frequency IHC rate responses are more consistent over clean and noisy conditions than state-of-the-art open-loop systems that neglect efferent feedback.

C. PHASE I – MODEL DEVELOPMENT

The overall goal of Phase I was to develop a closed-loop model of the auditory periphery that incorporates the human efferent system and to demonstrate the ability of such a model to produce displays of noisy speech that are more consistent with displays of speech in quiet than are displays produced by open-loop models. In embarking on this endeavor, we tested different models of cochlear filters, linear [Gammatone filters (Patterson, 1995)] as well as nonlinear [MBPNL (Goldstein, 1990)].

In implementing a cochlear model we use a bank of overlapping cochlear channels uniformly distributed along the ERB scale (Moore and Glasberg, 1983), four channels per ERB. Each cochlear channel comprises a filter (Gammatone or MBPNL) followed by a generic model of the IHC (half-wave rectification followed by a low-pass filter, representing the reduction of synchrony with CF). The dynamic range of the simulated IHC response is restricted – from below and above – to a "dynamic-range window" (DRW), representing the observed dynamic range at the AN level (i.e. the AN rate-intensity function); the lower bound and upper bound of the DRW stand for the spontaneous rate and rate-saturation, respectively.

C.1. Linear cochlear model with Gammatone filters

A linear Gammatone filter bank, which represents a linear based filtering strategy, was first examined as a baseline. Displays of the simulated IHC response were examined for noise intensity levels of 70, 60, and 50dbSPL and for SNR values of 20, 10, and 5dB. Figure A.2 provides a spectrographic example. The figure contains a 3-by-3 matrix of images; the abscissa represents the intensity of the background noise, in dB _SPL. The ordinate represents SNR, in dB. Each image represents the simulated IHC responses to the diphone s_a (duration of 249ms) spoken by a male speaker. Figure A.2 depicts the simulated open-loop Gammatone IHC response, with DRW=40dB. The position of the DRW was set such that speech is visible for the 50dbSPL×5dbSNR condition. Upper bound of the DRW was chosen such that 70dbSPL×20dbSNR condition is not oversaturated. A large inconsistency is observed across varying noise intensity and SNR levels. Note that for the DRW we chose, at 50dbSPL noise intensity level much of the speech energy is not present in the simulated IHC response for. Had the DRW range been shifted lower, more of the speech energy of the 50dbSPL noise intensity level would have been visible but also much noise.

C.2. Open-loop nonlinear cochlear model

A second model that we examined was Goldstein's Multi Band Pass Non Linear (MBPNL) model of nonlinear cochlear mechanics (Goldstein, 1990). This model operates in the time domain and changes its gain and bandwidth with changes in the input intensity, in accordance with observed physiological and psychophysical behavior. The MBPNL model is shown in figure A.3. The lower path (H1/H2) is a compressive nonlinear filter that represents the sensitive, narrowband compressive nonlinearity at the tip of the basilar membrane tuning curves. The upper path (H3/H2) is a linear filter (expanding function preceded by its inverse results in a unitary transformation) that represents the insensitive, broadband linear tail response of basilar-membrane tuning curves (after Goldstein, 1990). The parameter G controls the gain of the tip of the basilar membrane tuning curves, and is used to model the inhibitory efferent-induced response in the presence of noise (see Sec. C.3. below). For the open-loop MBPNL model the tip gain is set to G=40dB, to best mimic psychophysical tuning curves of a healthy cochlea in quiet (Goldstein, 1990).

The "iso-input" frequency response of an MBPNL filter at *CF* of 3400Hz is shown in figure A.4. The frequency response for the open-loop MBPNL model is shown at the upper-left corner (i.e. for G=40dB). For an input signal s(t)=$A\sin(\omega_0 t)$, with A and ω_0 fixed, the MBPNL behaves as a linear system with a fixed "operating point" on the expanding and compressive nonlinear curves, determined by A. Figure A.4 shows the iso-input frequency response of the system for different values of A. For a given A, a discrete "chirp" signal was presented to the MBPNL, with a slowly changing frequency. Changes in ω_0 occurred only after the system reached steady-state, for a proper gain measurement. For a 0dB input level A=1, the gain at *CF* is approximately 40dB. As the input level increases the gain drops and the bandwidth increases, in accordance with physiological and psycho-physical behavior.

Figure A.5 shows the simulated IHC response generated by the open-loop MBPNL to the diphone s_a (same as in Fig. A.2) for noise intensity levels of 70, 60, and 50dbSPL and for SNR values of 20, 10, and 5dB. The tip-gain is set to G=40dB and held constant for all SNR and noise levels. Here, we set DRW=22dB (down from 40dB for the Gammatone) because of the reduction in the overall dynamic range at the MBPNL output due to its inherent nonlinearity. The position of the DRW was chosen such that the speech energy of the simulated IHC response for the 70dbSPL×5dbSNR condition matched that of the same condition of the Gammatone model. Like the displays produced by the Gammatone model, the open-loop MBPNL displays show a large inconsistency across varying noise levels. Notice that for both open-loop models (Gammatone- and MBPNL- based) we could not find a "sweet-spot" for the DRW position that will provide a consistent display at the output, across rows and columns.

C.3. Cochlear model with efferent-inspired feedback

From the open-loop MBPNL model, we developed a closed-loop MBPNL model that includes an efferent-inspired feedback mechanism. Morphologically (e.g. Guinan, 1996), MOC neurons project to different places along the cochlea partition in a tonotopical manner, making synapse connections to the outer-hair cells and, hence, affecting the mechanical properties of the cochlea (e.g. increase in basilar-membrane stiffness). Therefore, we introduce a frequency dependent feedback mechanism which controls the tip-gain (G) of each MBPNL channel according to the intensity level of sustained noise at that frequency band. As shown in Fig. A.4, the upper-left panel represents the nominal response (i.e. healthy cochlea, in quiet), with the tip-gain G=40dB. By reducing G, the MBPNL response to weaker stimuli (e.g. background noise) is controlled. The lower right panel, for example, shows the MBPNL response for G=10dB. For high energy tone stimuli the MBPNL response is hardly affected, while the response for low energy stimuli (e.g. -80dB Re maximum input range) is reduced by some 30dB. In our efferent-inspired model, G is adjusted such that the average power of the cochlear output, in response to background noise at the input, will be such that the simulated IHC response to noise will be kept just below the lower bound of the DRW.

Figure A.6 depicts the simulated IHC response of an intermediate version of our closed-loop MBPNL model. DRW=22dB, its position is fixed at the same location as in the open-loop MBPNL model. The value of the tip gain (G) per cochlear channel is adjusted using the average power per frequency band, computed over 300ms duration of a speech-shaped noise preceding the speech signal. Due to the nature of the noise-responsive feedback, display of background noise is largely eliminated for all dbSPL×SNR conditions. At a given SNR, displays of processed noisy speech are consistent across db SPL noise level (rows in Fig. A.6). As expected, at a fixed dbSPL level, as the SNR drops (i.e. as the speech energy drops) the intensity of speech information in the spectrographic display dims (columns in Fig. A.6).

Figure A.8 shows the spectrographic displays of our current closed-loop MBPNL model, were the output of each MBPNL channel is normalized to a fixed dynamic range. The rational behind the normalization at the output stems from neurophysiological studies on anesthetized cats with noisy acoustic stimuli, which show that by stimulating the MOC nerve bundle electrically, the dynamic range of discharge rate at the AN is recovered (e.g. Winslow and Sachs, 1988)[5], as is illustrated in Fig. A.7. Upon visual examination, it can easily be seen that the displays are even more consistent across dbSPL×SNR conditions than those of Fig. A.6.

D. QUANTITATIVE EVALUATION

To obtain quantitative results, 96 processed noisy diphone pairs were compared in a simulated 2 alternative forced choice DRT test. Tests were run on the outputs of the open-loop Gammatone and the efferent-inspired closed-loop MBPNL models, after temporal smoothing. Template "states" were chosen for each DRT diphone-pair. In this study, the template states were the processed diphones at the 70dbSPL×10dbSNR condition (top two panels in figure A.9). The test stimuli were the same diphone tokens in different noise intensity levels and different values of SNR. For a given test token the MSE distance between the selected test token and the two

[5] Concurring with this observation are measurements of neural responses of awake cats to noisy acoustic stimuli, showing that the dynamic range of discharge rate at the AN level is hardly affected by changes in levels of background noise (May and Sachs, 1992).

template states was computed. The state template with the smaller MSE distance from the test token was selected as the simulated DRT response. Figure A.10 shows the average percent correct responses as a function of noise intensity level for the open-loop Gammatone (+) and the closed-loop MBPNL (×). Average is over all DRT words and all SNR values. As the plot indicates, the closed-loop MBPNL model behaved more consistently over all noise intensity levels than the open-loop system. The performance of the open-loop system significantly degraded as the noise intensity level varied further from the template noise intensity level (70dbSPL in this example). Figure A.11 shows a more detailed version of Fig. A.10; errors – averaged is over all DRT words – are plotted as a function of SNR, with noise intensity (in dbSPL) as a parameter. For the open-loop model best performance occurs at 70dbSPL – the template noise condition (as expected, no errors occur at 10dbSNR – the template SNR). The extent of inconsistency is reflected by the poor (close to chance) performance at all other noise intensities, for all SNR values (an unexplained exception is the 60dbSPL×20dbSNR condition). In contrast, performance with the closed-loop MBPNL model is very consistent across all conditions. Figure A.12 is yet another way of looking at the same data; here, errors are plotted as a function of noise intensity, with SNR as the parameter. Similar conclusion can be drawn, i.e. for the open-loop model, for each SNR best performance occurs at 70dbSPL (the template noise condition); at all other noise intensity levels performance is close to chance. Far fewer errors are made when the closed-loop model is used; most the errors are in noise intensity levels away from the template noise condition.

E. SUMMARY

This report summarizes work that has been completed in Phase I of the STTR program. We have implemented a closed-loop model of the auditory periphery with an efferent-inspired feedback and have quantitatively demonstrated its ability to produce spectrograms of noisy speech samples that are far more consistent with spectrograms of speech in quiet than are spectrograms produced by an open-loop model of the auditory periphery. This increase in performance in noise and increased robustness mimics the general observed behavior of humans. Whether this model of auditory periphery preserves phonetic information in patterns that follow psychophysical patterns will be rigorously inspected during Phase II, where the central part of the proposal, i.e. the formulation of a perception-based distance measure, will be established.

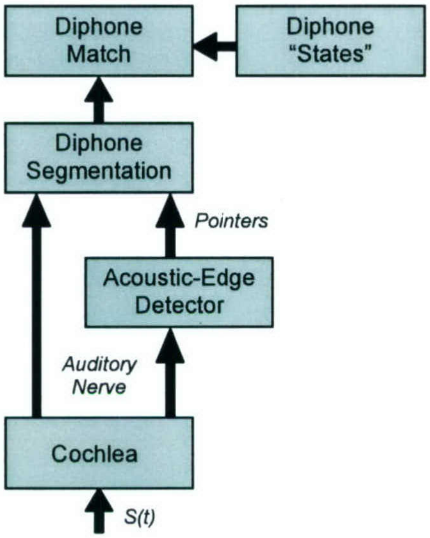

Figure A.1. A schematic description of our conceptual model of perception of diphones

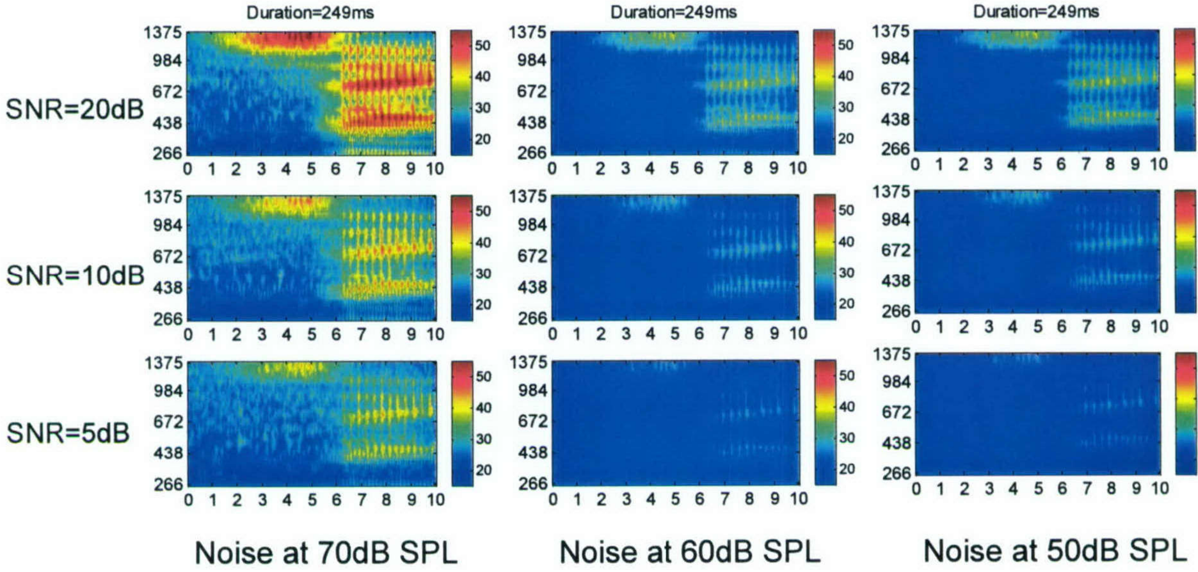

Figure A.2. Simulated IHC response to diphone s_a, produced by an open-loop Gammatone model; DRW=40db; Position of DRW set such that speech is visible for the 50 db SPL Noise and SNR=5db condition. Upper bound of DRW chosen such that 70dB_SPL×SNR=20dB condition is not oversaturated.

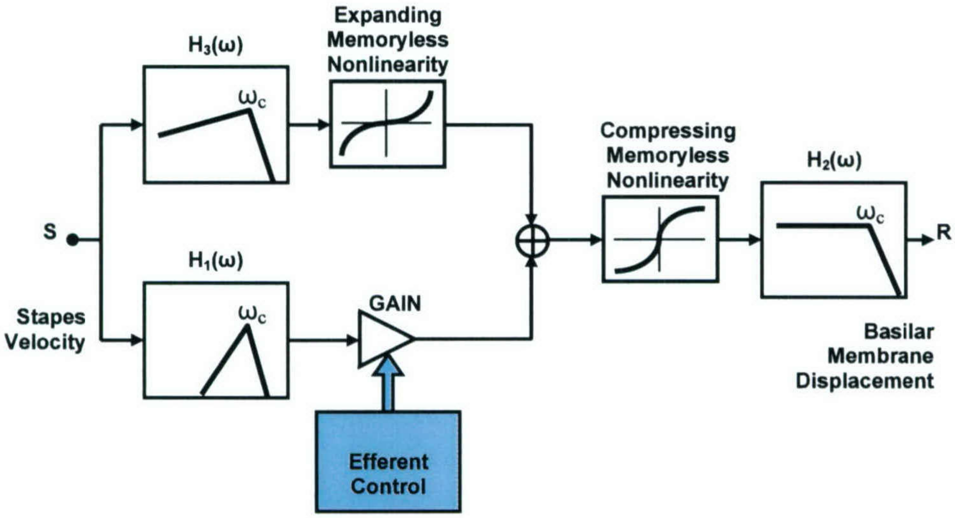

Figure A.3. Goldstein's MBPNL model

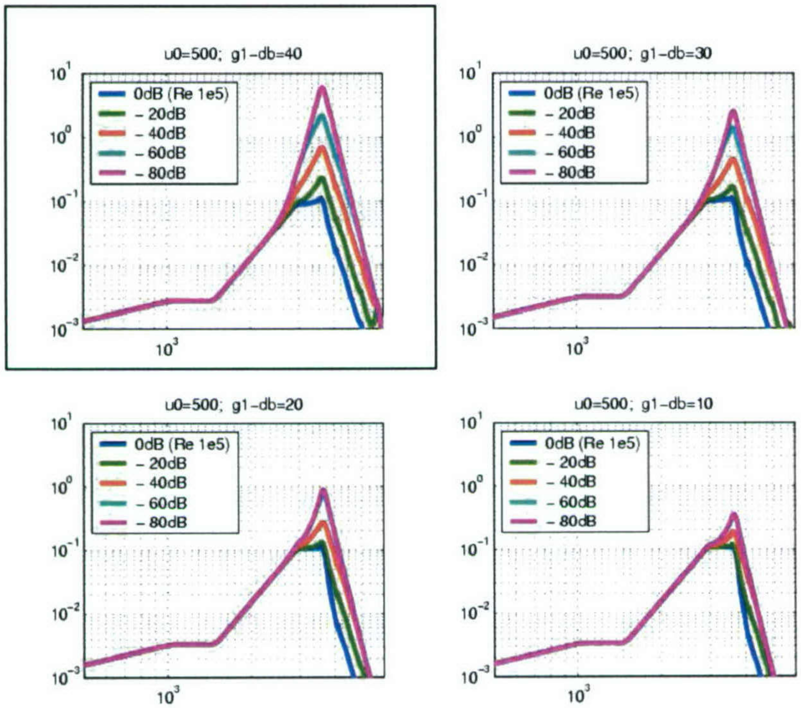

Figure A.4. Iso-input frequency responses of an MBPNL filter (at *CF* of 3400Hz) for different values of tip-gain, G. From Upper-left, clockwise: G=40, 30, 20 and 10dB. Upper-left corner (G=40dB) is for healthy cochlea in quiet (Goldstein, 1990).

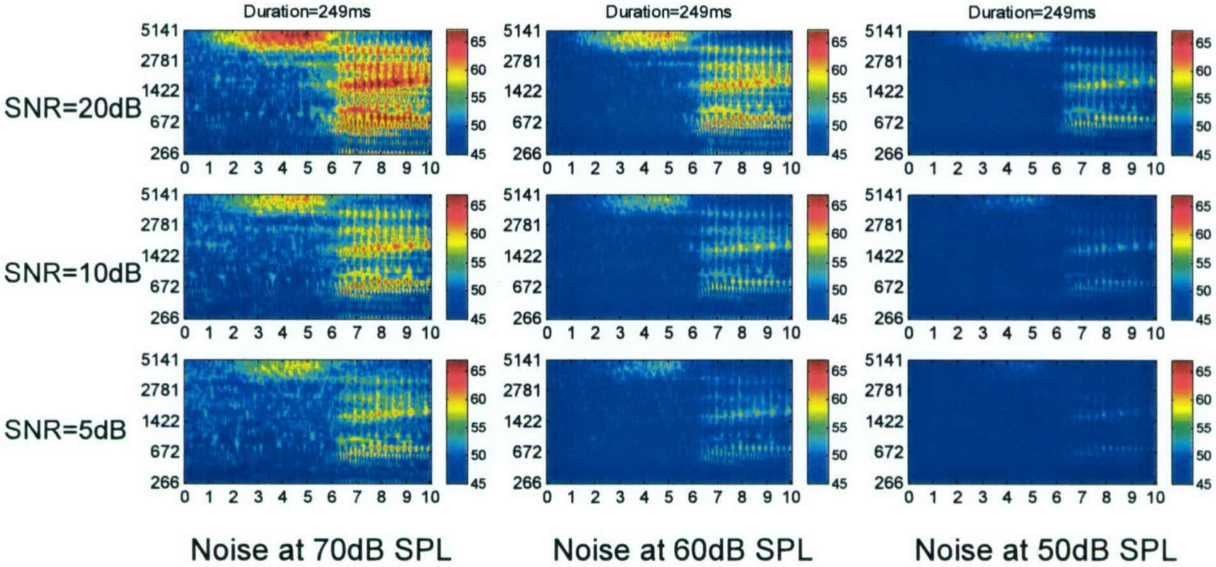

Figure A.5. Simulated IHC response to diphone s_a, produced by an open-loop MBPNL model; Fixed G=40dB; DRW=22dB; DRW chosen to approximately match speech power of the Open loop Gammatone model displays of figure 2.

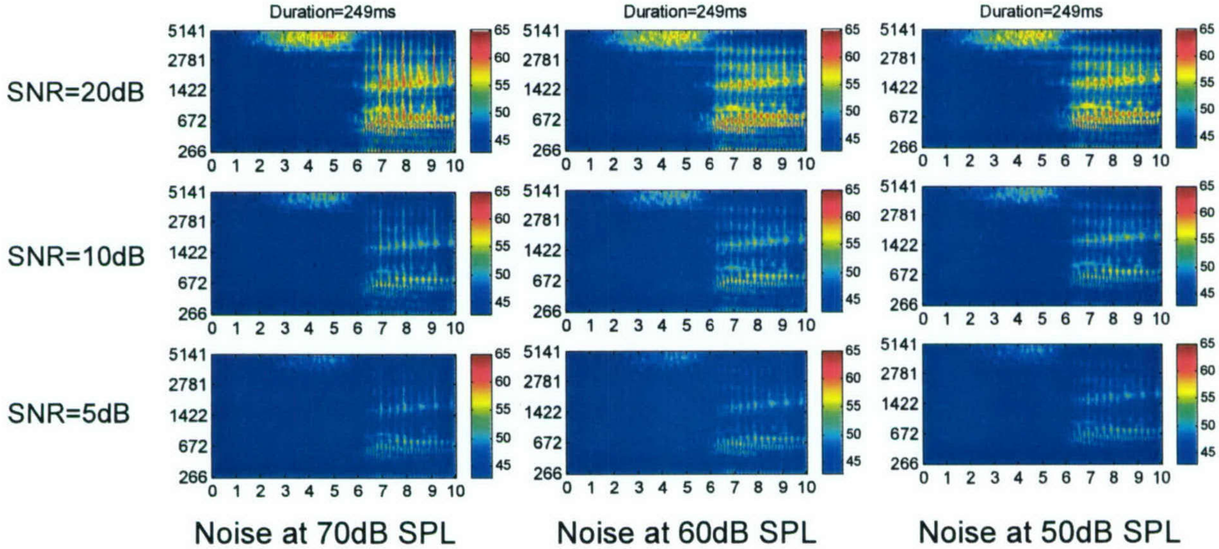

Figure A.6. Simulated IHC response to diphone s_a, produced by an intermediate closed-loop MBPNL model. DRW is same as in open-loop MBPNL mode.

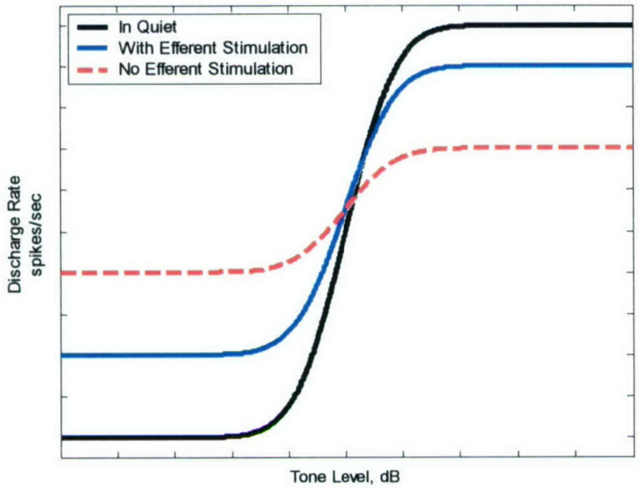

Figure A.7. Illustration of the observed efferent-induced dynamic range recovery of the discharge rate in the presence of background noise (e.g. Winslow and Sachs, 1988). Discharge rate versus Tone level is cartooned in quiet condition (full dynamic range, black); anesthesized cat, i.e. no efferents activity (much reduced dynamic range, red) and with electrical stimulation of COCB nerve bundle.

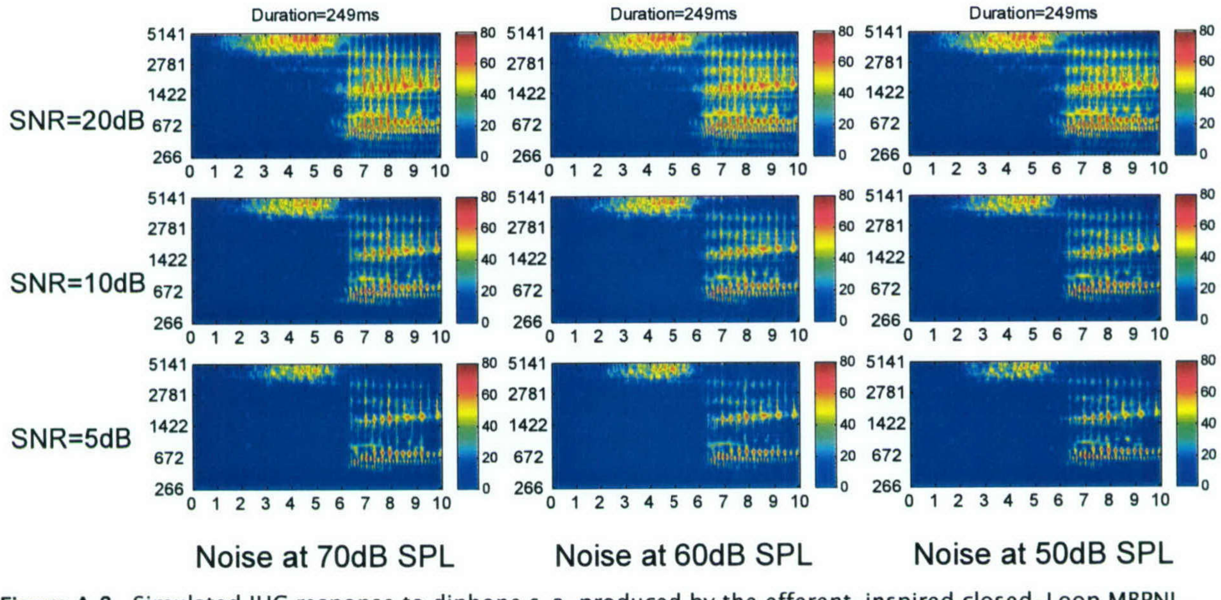

Figure A.8. Simulated IHC response to diphone s_a, produced by the efferent-inspired closed-Loop MBPNL. DRW is same as in open-loop MBPNL mode. Output of each MBPNL channel is normalized to a fixed dynamic range.

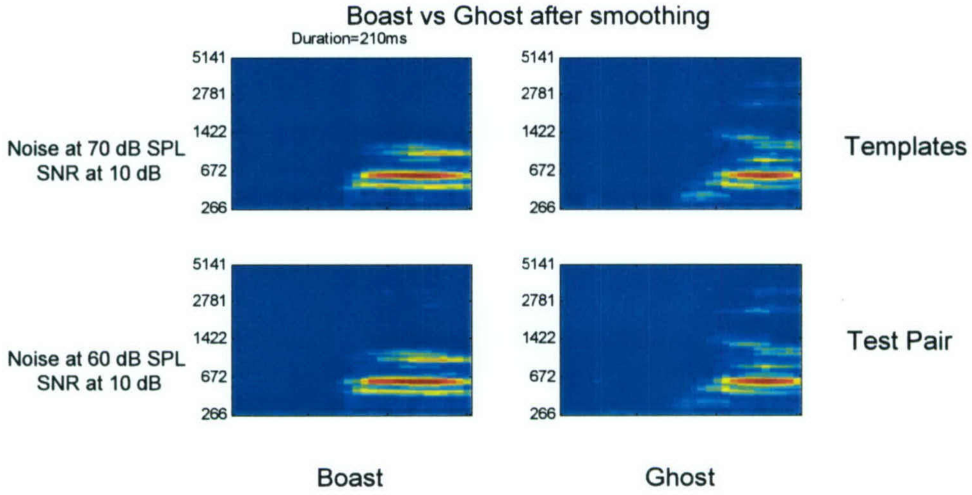

Figure A.9. Temporally smoothed simulated IHC response produced by the efferent-inspired closed-Loop MBPNL (with normalization at the output). Representations at the 70dB_SPL×SNR=10dB condition are chosen as template "states". A mimic of the "one-interval two-alternative forced-choice" paradigm is conducted for each DRT word-pair.

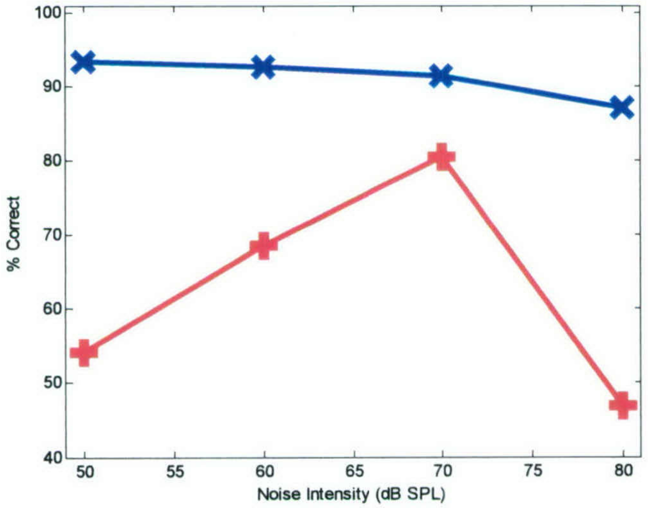

Figure A.10. Percent correct responses as a function of noise intensity level for the open-loop Gammatone (+) and the closed-loop MBPNL (×), using the 70dB_SPL×SNR=10dB condition as template. Average is over all DRT words and all SNR values.

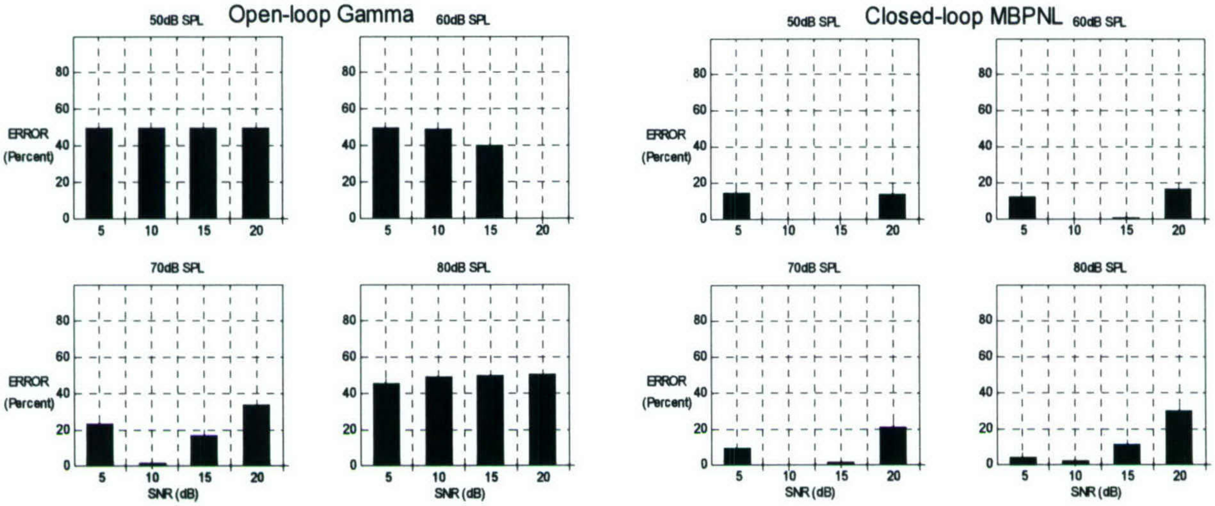

Figure A.11. Same data as in Fig. A.10, in more details. Errors (in percent) are averaged over all DRT words and plotted as a function of SNR, with noise intensity (in dB_SPL) as a parameter

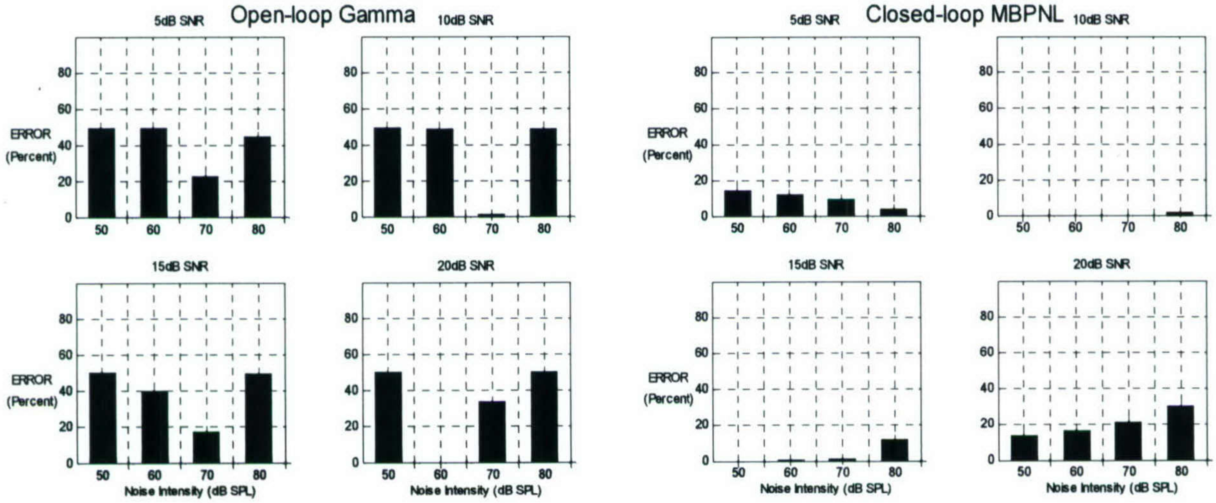

Figure A.12. Same data as in Fig. A.10, in more details. Errors (in percent) are averaged over all DRT words and plotted as a function of noise intensity, with SNR as a parameter.

BIBLIOGRAPHY (Appendix A)

Dewson, J. H. (1968). Efferent olivocochlear bundle: some relationships to stimulus discrimination in noise. *J. Neurophysiol.*, 31:122–130.

Ghitza, O. (2004). On the possible role of MOC efferents in speech reception in noise. JASA, vol. 115(5), abst., page 2500.

Giraud, A. L., Garnier, S., Micheyl, C., Lina, G., Chays, A., and Chery-Croze, S. (1997). Auditory efferents involved in speech-in-noise intelligibility. *NeuroReport*, 8:1799-1783.

Goldstein, J. L. (1990). Modeling rapid waveform compression on the basilar membrane as a multiple-bandpass-nonlinearity filtering, *Hear. Res.*, **49**, 39-60.

Guinan, J. J. (1996). Physiology of Olivocochlear Efferents. In Dallos, P., Popper, A. N. and Fay, R. R., editors, *The Cochlea*, pages 435–502, Springer, New-York.

May, B. J. and Sachs, M. B. (1992). Dynamic range of neural rate responses in the ventral cochlear nucleus of awake cats. *J. Neurophysiol.*, 68:1589–1603.

Moore, B. C. J. and Glasberg, B. R. (1983). Suggested formula for calculating auditory-filter bandwidth and excitation patterns, *J. Acoust. Soc. Am.*, **74,** 750-753.

Patterson R. D., Allerhand M. H., and Giguere C. (1995). Time-domain modeling of peripheral auditory processing: a modular architecture and a software platform. *J. Acoust. Soc. Am.*, **98**, 1890-4.

Voiers, W. D. (1983). Evaluating processed speech using the diagnostic rhyme test. *Speech Technology*, 1(4):30–39.

Winslow, R. L. and Sachs, M. B. (1988). Single-tone intensity discrimination based on auditory-nerve rate responses in backgrounds of quiet, noise, and with stimulation of the crossed olivocochlear bundle. *Hear. Res.*, 35:165–190.

Zeng, P. G., Martino, K. M., Linthcum, F. H., and Soli, S. (2000). Auditory perception in vestibular neurectomy subjects. *Hear. Res.*, 142:102–112.

Appendix B

A Brief Overview of Voiers' DRT

The DRT (Diagnostic Rhyme Test) version of Voiers (1983) is a way of measuring the intelligibility of processed speech and has been used extensively in evaluating speech coders. From an acoustic point of view, Voiers' DRT database covers initial dyads of spoken CVCs. The database consists of 96 pairs of confusable words spoken in isolation. Words in a pair differ only in their initial consonants. The dyads are equally distributed among 6 acoustic-phonetic distinctive features and among 8 vowels (hence 2 word-pairs per [quadrant×feature] cell). The feature classification (outlined in Table 1) follows the binary system suggested by Jakobson, Fant and Halle (Jakobson *et al*., 1952), and the vowels are [ee] and [it] (High-Front), [eh] and [at] (High-Back), [oo] and [oh] (Low-Front) and [aw] and [ah] (Low-Back). In our version of the DRT the vowels are collapsed into 4 quadrants (High-Front, High-Back, Low-Front, Low-Back), hence 4 word-pairs per a [quadrant×feature] cell.

The psychophysical procedure is carefully controlled to assure a task with minimum cognitive load. The listeners are well trained and are very familiar with the database, including the voice quality of the individual speakers. The experiment uses a one-interval two-alternative forced-choice paradigm. First, the subject is presented visually with a pair of rhymed words. Then, one word of the pair (selected at random) is presented aurally and the subject is required to indicate which of the two words was played. This procedure is repeated until all the words in the database have been presented. In our version of the DRT words are played sequentially, one every 2.5 – 3 seconds; the visual presentation precedes the aural presentation by 1sec., and the decision (binary) must be made within 1sec of the aural presentation. Words in the database are divided into “runs”, and the duration of one run is limited to about 2.5 minutes (to avoid fatigue).

The scores of one complete DRT-session will be tabulated with a cell granularity of [quadrant×feature], as illustrated in Table 2. A table-entry contains the number of words per cell that where mistakenly identified; it is an integer between 0 and 4, since the total number of words per cell is 4.

Our knowledge about the acoustic correlates of the Jakobsonian dimensions provides diagnostic information about temporal representation of speech, while the vowel quadrant identity provides information about the frequency range (i.e. location of the formants in action). Hence, the integrated information can link phonetic confusions with their origin in the time-frequency plane. We shall utilize the usage of such linkage to guide the procedure of tuning the parameters of the auditory model.

Table 1. Samples of word-pairs used in Voiers' DRT (1983).

Voicing (VC) (*Voiced - Unvoiced*)	**Nasality (NS)** (*Nasal - Oral*)	**Sustention (ST)** (*Sustained -Interrupted*)
veal - feel zed - said –	meat - beat neck - deck –	vee - bee fence - pence –
Sibilation (SB) (*Sibilated - Assibilated*)	**Graveness (GV)** (*Grave - Acute*)	**Compactness (CM)** (*Compact - Diffuse*)
cheep - keep jot - got –	peak - teak wad - rod –	key - tea got - dot –

Table 2. A sample of the outcome of one DRT session, one stimulus condition, and one subject. A table-entry contains the number of words per [quadrant×feature] bin mistakenly identified (an integer between 0 and 4).

	VC		NS		ST		SB		GV		CM	
	+	–	+	–	+	–	+	–	+	–	+	–
High-Front	0	0	1	1	0	4	2	2	2	1	1	1
High-Back	1	1	2	0	2	1	1	0	1	3	0	0
Low-Front	1	0	0	0	1	3	0	1	1	4	1	1
Low-Back	1	1	1	1	3	4	2	3	3	2	1	0

Table 3. The Jakobsonian dimensions and their acoustical correlates

Voicing	- Periodicity and shorter time of onset duration (Voiced) - Discriminability - at [0,1000] Hz
Nasality	- Formants at 200, 800 and 2200~Hz - Nulls throughout the frequency range (Nasals) - Discriminability - at [0,1000] Hz
Sustention	- Gradual onset and presence of mid-frequency noise (Sustained) - Durational and high-frequency cues
Sibilation	- Higher-frequency noise and greater duration (Sibilant) - Duration is most important acoustical correlate
Graveness	- Origin and direction of second-formant transitions - Grave consonants - steep upward transitions - Acute consonants - downward second-formant transitions - Greater concentration of low-frequency energy (Grave)
Compactness	- Concentration of spectral energy at mid-frequency range (Compact) - More-widely separated spectral peaks (Diffused)

BIBLIOGRAPHY (Appendix B)

Voiers, W. D. (1983). Evaluating processed speech using the diagnostic rhyme test. *Speech Technology*, 1(4):30–39.

Jakobson, R., Fant, C. G. M., and Halle, M. (1952). Preliminaries to speech analysis: the distinctive features and their correlates. Technical report, Acoustic Laboratory, Massachusetts Institute of Technology.